Analysis of domain correlations in yeast protein complexes

Doron Betel\textsuperscript{1,2}, Ruth Isserlin\textsuperscript{1,2} and Christopher W. V. Hogue\textsuperscript{1,2,*}

\textsuperscript{1}Samuel Lunenfeld Research Institute, Mt Sinai Hospital, Toronto, M5G 1X5, Canada and \textsuperscript{2}Department of Biochemistry, University of Toronto, Toronto, MSS 1A8, Canada

Received on January 15, 2004; accepted on March 1, 2004

ABSTRACT

Motivation: A growing body of research has concentrated on the identification and definition of conserved sequence motifs. It is widely recognized that these conserved sequence and structural units often mediate protein functions and interactions. The continuing advancements in high-throughput experiments necessitate the development of computational methods to critically assess the results. In this work, we analyzed high-throughput protein complexes using the domain composition of their protein constituents. Domains that mediate similar or related functions may consistently co-occur in protein complexes.

Results: We analyzed \textit{Saccharomyces cerevisiae} protein complexes from curated and high-throughput experimental datasets to identify statistically significant functional associations between domains. The resulting correlations are represented as domain networks that form the basis of comparison between the datasets, as well as to binary protein interactions. The results show that the curated datasets produce domain networks that map to known biological assemblies, such as ribosome, RNA polymerase, proteasome regulators, transcription initiation and histones. Furthermore, many of these domain correlations were also found in binary protein interactions. In contrast, the high-throughput datasets contain one large network of domain associations. High connectivity of RNA processing and binding domains in the high-throughput datasets reflects the abundance of RNA binding proteins in yeast, in agreement with a previous report that identified a nucleolar protein cluster, possibly mediated by rRNA, from these complexes.

Availability: The software is available upon request from the authors and is dependent on the NCBI C++ toolkit.

Contact: hogue@mshri.on.ca

INTRODUCTION

Advances in experimental techniques enable researchers to conduct large-scale experiments that survey many proteins in a short time span. In the recent years, we have witnessed an unprecedented increase in protein interaction data generated by high-throughput experiments, such as, yeast two-hybrid and protein complex identification experiments (Phizicky \textit{et al.}, 2003; Tyers and Mann, 2003). Inherent to these methodologies is a high ratio of false positive results that require post-processing and careful analysis of the data. Previous analyses of high-throughput protein complex data were generally performed on the protein constituents of the complexes and their functional classifications as assigned by ontology consortiums [Gene Ontology (GO) and MIPS] (Bader and Hogue, 2002; von Mering \textit{et al.}, 2002). Some of the conclusions from these studies point to the lack of internal consistency and a high ratio of erroneous data. Notwithstanding, there remains a considerable amount of insight that can be extracted from these datasets by examining other types of protein properties, such as co-expression patterns and deletion phenotypes (Dezso \textit{et al.}, 2003).

Our approach to the analysis of high-throughput protein complexes stems from the notion that the biological role of a complex is determined by the domain composition of its protein constituents. A corollary of the above thesis is that domain associations derived from protein complexes will reflect the biological role of these complexes. By examining the domain composition of protein complexes from curated datasets and high-throughput experiments, we were able to assess the functional information that is captured in these associations as well as to identify experimental bias.

In this work, we introduce a novel approach to studying protein complexes from different experimental sources. We begin by identifying statistically significant domain pairs that co-occur in protein complexes. These domain correlations are generated from high-quality curated protein complexes as well as from high-throughput experimental datasets.

The advantages of using the domain composition as the basis of comparison include enrichment of functional information and eliminating the need to model internal complex connectivity. Different proteins with similar activities are likely to contain similar domains. As a result, there are multiple instances of the domains dispersed in the datasets, whereas there might be only a few copies of each protein.
This provides an amplification of the functional information content. Additionally, the correlations between domains with different functions might demonstrate connections between the biological processes that these domains facilitate. Previous comparison of protein complexes with protein interactions required some initial assumption about the internal topology of the complexes. Protein complexes were typically represented as ‘spoke’ or ‘matrix’ models of binary interactions to provide a common unit of comparison to evaluate protein complexes to interactions (Bader and Hogue, 2002). Domain correlations can represent functional coupling of domains in complexes or binary interactions. Therefore, the need to model interactions within complexes is eliminated and yet complexes are comparable with binary protein interactions.

In this analysis, we compare the datasets using their domain correlation networks and highlight some of the major differences in their domain composition, biological content and network topology. We investigate further how these domain correlations were compared with the domain composition in binary protein interactions. By identifying correlated domain pairs that are supported by protein interactions, we are able to compare the protein complexes data to binary protein interactions on the basis of the underlying biological functionality represented in each dataset. This method enables us to highlight some biologically significant domain correlations that can be demonstrated in known binary protein interaction.

CONSTRUCTION

Datasets and domain annotation

Two recent large-scale proteomic studies have identified a large number of yeast molecular complexes using mass spectrometry techniques—referred hereafter as TAP (Gavin et al., 2002) and HMS-PCI (Ho et al., 2002) datasets. The principle methodology in both studies is based on tagging a large number of proteins of interest (baits) and using them to purify whole complexes from the cell with minimal disruption of the complexes. In addition to these two sources, we used 235 high-quality curated complexes from the MIPS Yeast Genome Database (Mewes et al., 2002) downloaded in February 2002 and 168 complexes from the Yeast Proteome Database (YPD) from March 2001 (Costanzo et al., 2001). We developed a computational platform that provides quality domain annotation integrating CDD (Marchler-Bauer et al., 2003) and InterPro (Mulder et al., 2003) resulting in a unique set of InterPro domains for every protein and mapping of domain entries to GO annotation (Camon et al., 2003).

Definition of domain correlation

A domain–domain correlation is defined as the co-occurrence of two domains from two different proteins in the same molecular complex with significant probability. The number of co-occurrences for each domain pair is summed up over the entire collection of complexes. In simple words, we count the number of times two domains appear on different proteins in the same complex over the entire set of complexes. A more formal definition is provided in the footnote to Table 1.

Statistical sampling

Computing probability values for domain co-occurrences requires a distribution of random measurements that reflect the background co-occurrences between domains in the absence of any biological driving force (i.e. by chance alone). The observed number of domain co-occurrences can then be compared with this random model. We devised two methods to generate these control sets. The first is a domain randomization scheme in which domains are randomly redistributed between the proteins in the datasets. The second is a protein randomization scheme in which the proteins are randomly redistributed among the complexes while keeping

### Table 1. Summary of the domain correlation networks from the four datasets—YPD (Costanzo et al., 2001), MIPS (Mewes et al., 2002), TAP (Gavin et al., 2002) and HMS-PCI (Ho et al., 2002)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of complexes</th>
<th>Number of proteins (with domains)</th>
<th>Number of single-domain proteins</th>
<th>Number of correlated domains (vertices)</th>
<th>Number of correlations (edges)</th>
<th>Number of network components</th>
</tr>
</thead>
<tbody>
<tr>
<td>YPD</td>
<td>168</td>
<td>257</td>
<td>173</td>
<td>626/606/506</td>
<td>755/735/485</td>
<td>209/239/259</td>
</tr>
<tr>
<td>TAP</td>
<td>232</td>
<td>643</td>
<td>461</td>
<td>1377/1387/1135</td>
<td>249/321/142</td>
<td>66/7/18</td>
</tr>
<tr>
<td>HMS-PCI</td>
<td>511</td>
<td>983</td>
<td>657</td>
<td>1666/1533/1333</td>
<td>369/368/192</td>
<td>43/3/20</td>
</tr>
</tbody>
</table>

For each dataset, the number of domain correlations are reported when using the domain (a) and protein (b) randomization schemes, as well as the overlap between the two sets (c). The number of components indicates the number of independent domain components (unconnected sub-networks), which the network is composed of. Formally we define the set, C, of all domain co-occurrences (dx, dy) in a complex as Sc = {(dx, dy) | dx ∈ Pj, dy ∈ Pi} for all Pi ∈ C, Pj ∈ C and i ≠ j. That is, Pi and Pj are two different proteins in complex C, dx, dy are domains in Pi and Pj, respectively. Similarly, the set, Sc, of all domain correlations in the entire dataset is Sc = ∪Pi∈D, where D is the collection of complexes in the dataset and C is a complex in D. The number of co-occurrences between any two domains, x and y, is the sum of all the (dx, dy) elements in the set, Sc.

*a Using domain randomization scheme.

*b Using protein randomization scheme.

C Overlap between a and b.
their correct domain annotation. For each of the four datasets we performed 3000–5000 randomization cycles using the two randomization methods such that every domain pair had a large number of random co-occurrences. Comparing the random co-occurrences with the observed co-occurrences we computed two P-values, one from each randomization method, for every domain pair. The P-value reflects the statistical significance of observing the two domains in the same molecular complex.

It is intuitive to represent the domain correlations as a graph where every vertex represents a domain and every edge corresponds to an association between two domains with the edge weight equal to the P-value for that association. For each of the four datasets we generated two networks; one derived from the domain randomization P-values, and a second from the protein randomization P-values. Finally, a third network representing domain correlations that were found to be significant in both types of randomization was generated by the intersection of the above two networks and was used in the final analysis.

In each of the four datasets there is a considerable overlap between the two network graphs generated from the domain and protein randomizations (34–63% common edges). This is largely due to the fact that the majority of the proteins contain a single domain (67–71%, Table 1). The two randomization methods are similar in the sense that randomizing proteins in complexes is the same as randomizing domains in proteins when applied to single-domain proteins. Additionally, the overlap suggests that in most cases the correlations of any given domain are tightly coupled with the functional association of its neighbouring domains in the same protein. Hence, it is expected that domains residing on the same protein will form similar functional relationship with other domains.

**ANALYSIS**

**Global network topologies**

The graphs generated from the YPD and MIPS datasets (Fig. 1) contain independent sub-networks of domain correlations corresponding to well-defined biological processes. The two largest sub-networks in the MIPS graphs (Fig. 1A) correspond to the small and large ribosomal subunits (note that ribosomal proteins were excluded from both the TAP and HMS-PCI complexes). Other biological processes are also well represented, such as the histone group, cyclin-dependent kinase, RNA polymerase, translation and a number of transcription factor complexes. Biological process classifications of domains by GO terms from InterPro (Camon et al., 2003) further verifies the clustering of domains with related functionality. However, not all domains are annotated with GO terms and some contain multiple classifications.

In contrast to the curated data, the high-throughput datasets produce large networks with a few central hubs linking all the domains, typical of small-world networks (Strogatz, 2001). Among the most highly connected domains in the HMS-PCI set (Fig. 2B) are the kinases and cell signalling modules. This is an expected bias from the bait selection that included 100 protein kinases, which is significantly different from the TAP bait selection (Ho et al., 2002). Surprisingly, DEAD (DEAD/DEAH box helicase, InterPro ID IPR001410), Helicase_C (Helicase, C-terminal, IPR001650) and RNA_rec_mot (RRM, IPR000504) domains, in both TAP and HMS-PCI networks (Fig. 2), are ranked among the most correlated domains in the networks and all three are implicated in RNA processing. Helicase_C is found in many helicase proteins, RRM is an RNA recognition motif found in a variety of RNA binding proteins (Bandziulis et al., 1989) and DEAD represents a large family of ATPases involved in nuclear pre-mRNA splicing, ribosome assembly and RNA degradation (Schmid and Linder, 1992).

False positive protein identification in high-throughput experiments inevitably affects any analysis performed thereafter. Therefore, fortuitous proteins in complexes may influence this domain analysis by introducing erroneous domain correlations. The result being that the HMS-PCI and TAP network do not separate into individual biological complexes as compared with the high-quality curated datasets.

Previous analysis of protein interactions point to the scale-free nature of these networks and suggest that proteins, which play key role in cellular process are probably centrally connected (Jeong et al., 2001). Results from the current domain correlation are not necessarily congruent with these conclusions. The scale-free nature of the networks cannot be conclusively determined due to the small number of nodes. Although a number of domains in the high-throughput sets appear to be hubs, they reflect the abundance of particular functional processes (such as RNA processing) and are not necessarily elements of essential proteins.

**Supported domain correlations**

A natural question that arises from the domain correlations is: To what degree are these domain relationships represented in other types of protein interactions data? To investigate this question, we collected a non-redundant set of interactions from BIND (Bader et al., 2003), DIP (Xenarios et al., 2001) and the literature [through the use of PreBIND (Donaldson et al., 2003)] excluding all interactions that were derived from protein complex identification experiments. We identified all correlated domain pairs that are supported in this collection of interactions. A domain correlation is considered corroborated if there are two interacting proteins, each containing one of the correlated domains.

A breakdown of the supported domain pairs by datasets (Table 2) shows that the curated datasets contain a higher percentage of validated correlations. The small proportion of supported domain pairs in the MIPS dataset is misleading since a large portion of the MIPS network comprises ribosomal motifs. Indeed, without the ribosomal domain
Fig. 1. Domain correlation graphs for the MIPS (Mewes et al., 2002) (A) and YPD (Costanzo et al., 2001) (B) datasets. A total of 168 complexes from YPD and 260 from MIPS database were analyzed for statistically significant domain co-occurrences. Each vertex represents a domain and every edge corresponds to a significant correlation between domains ($P$-value < 0.05). Domains are coloured by GO biological function annotation as assigned by InterPro (Camon et al., 2003) and labels correspond to InterPro short names. Edges are coloured according to the source of supporting interactions: orange, unsupported; blue, supported by either BIND or DIP; green, supported by PreBIND yeast Medline search; and brown, supported by BIND/DIP + PreBIND yeast abstracts. A number of network components correspond to well-known biological complexes including the ribosome and RNA polymerase, whereas others represent general biological assemblies, such as cytoskeletal formation, elongation factor, DNA replication initiation, mismatch repair and transcription activation. Shaded areas are specific correlations discussed in the text.
Fig. 2. The TAP (Gavin et al., 2002) (A) and HMS-PCI (Ho et al., 2002) (B) domain correlation networks. Edge and vertex colouring are similar to Figure 1. Both networks contain one large sub-graph that include ∼80% of the domains in each dataset. Note that in both networks the RNA processing domains (the ATPase DEAD, C-terminal helicase and RNA recognition motif) are highly connected. The HMS-PCI bait selection preference is reflected in the network by the high connectivity of the kinase domains.
D. Betel et al.

**Table 2.** The ratio of domain correlations that are supported by binary protein interaction data from BIND (Bader et al., 2003) or DIP (Xenarios et al., 2001) or PreBIND yeast (Donaldson et al., 2003) to the total number of correlations

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Edges from domain randomization</th>
<th>Edges from protein randomization</th>
<th>Overlap</th>
</tr>
</thead>
<tbody>
<tr>
<td>YPD</td>
<td>48/75 (64%)</td>
<td>53/77 (69%)</td>
<td>33/48 (69%)</td>
</tr>
<tr>
<td>MIPS</td>
<td>89/572 (15.5%)</td>
<td>106/586 (18%)</td>
<td>51/290 (17.5%)</td>
</tr>
<tr>
<td>TAP</td>
<td>46/249 (18.5%)</td>
<td>85/321 (26.5%)</td>
<td>33/142 (23%)</td>
</tr>
<tr>
<td>HMS-PCI</td>
<td>69/369 (19%)</td>
<td>104/568 (18%)</td>
<td>35/192 (18%)</td>
</tr>
</tbody>
</table>

We excluded all interactions derived from protein complexes. Shown are the ratios from the domain randomization, protein randomization schemes and overlap of the two. Note that when excluding ribosomal correlations from the MIPS overlap network the ratio of supported correlations is 46%.

correlations 46% of the MIPS correlations (data not shown) and 69% of the YPD correlations are supported by protein interactions. Internal ribosomal protein interactions are typically ignored in interaction experiments. However, since the ribosome is a large molecular complex it is reasonable to presume that many of the ribosomal domain correlations would be represented if the databases contained the interactions of the ribosome complex. Within the high-throughput datasets, the TAP domain correlations have 5% more supported correlations than the HMS-PCI following previous conclusion that the TAP dataset has more literature-supported interactions (Bader and Hogue, 2002). It is interesting to note that although only ~20% of the domain correlations from high-throughput complexes are directly supported by protein interactions, this is a larger overlap than that shown by the direct comparison of protein complexes to binary interactions (Bader and Hogue, 2002; von Mering et al., 2002). The direct comparison of protein complexes to protein interaction requires some initial assumption about the internal protein interactions within complexes. The current comparison suggests that although these experiments failed to reproduce many of the known interactions they were able to capture more biologically relevant relationships than those previously assumed. The supported domain correlations are from a small set of biological processes that are well represented in the yeast protein interaction data. The majority include kinase and cell signalling domains, proteasome activity, RNA processing, cytoskeletal rearrangements and transcription. These results imply that most of the binary protein interaction experiments (either high-throughput or directed) have so far sampled subsets of the functional relations present in the yeast proteome.

**BIOLOGICAL INSIGHT**

**Global differences**

The apparent differences in the topology of the domain correlation networks from different datasets, as well as the degree of overlap with binary interaction data, underscore the basic differences between protein complexes that were detected by bait pull-down experiments and those identified by multiple experiments. The molecular complexes in the YPD and MIPS datasets contain protein assemblies with distinct biochemical activities, such as the ribosome, RNA polymerase, mismatch repair, proteasome and transcription factors. Complexes in these datasets are the result of directed experiments that observe a specific subset of proteins. These complexes are compiled from heterogeneous proteins with stable physical interactions that work in unison to perform the function of the complex. In contrast, the high-throughput data are a collection of associated proteins that were identified by a few undirected experiments; however, biased by bait selection and purification method. These complexes may contain transient, loosely connected proteins that could be linked to each other through diverse continual biological events, such as signal pathways, transcription and molecular transport. A resulting complex is not made from a collection of heterogeneous sub-units belonging to a solid complex but rather represents a snapshot of the molecular associations the bait protein has at the time of purification. Therefore, a pull-down complex represents a population of different biological complexes with which the bait protein is associated. It is not likely that high-throughput experiments will produce domain correlation networks that recapture literature-based functional complexes like those seen in the curated datasets.

**Nucleolar**

Protein interaction graphs often contain regions of high local density where the proteins in the region are preferentially connected to each other while sparsely linked to proteins outside the dense region. A previous publication using a k-core network extraction algorithm, which detects local dense regions (Bader and Hogue, 2002), identified the components of the nucleolar protein complex by combining the TAP and HMS-PCI datasets. In the nucleolar complex, rRNA is transcribed and packaged with other components of the ribosome mechanism. Interestingly, many of the proteins in that complex are implicated in RNA processing and contain the DEAD, Helicase_C and RNA_rec_mot domains, which were identified as highly connected hubs in the TAP and HMS-PCI domain correlation networks (Fig. 2).

The majority of the complexes in the high-throughput studies contained proteins of high abundance involved in RNA translation and ribosomal components that were extracted alongside the bait’s natural associations. Both the HMS-PCI and TAP complexes were not treated with RNase but the ribosomal proteins were filtered out in the final analysis (Gavin et al., 2002; Ho et al., 2002). It is known from microscopic studies of the nucleolus that the fibrillar, dense fibrillar and granular regions of the nucleolus are interconnected by rRNA (Olson et al., 2000). The lack of RNase treatment and enrichment of RNA processing proteins imply that rRNA may have bridged a significant portion of the protein links...
in the high-throughput complexes and hence are not direct protein–protein interactions as first suggested.

**Biological examples**

Most of the domain correlations in the curated datasets can be traced to known biological complexes such as the cyclin-kinase group, histone group and proteasome regulation. A more specific example is demonstrated by the binding of Cdc45 protein to MCM5(Cdc46) that is required for the initiation of DNA replication in yeast. This can be generalized as the functional coupling between CDC45 domain (IPR003874) and MCM family of proteins (IPR001208), which are both required for DNA replication (Saha et al., 1998). Indeed, the MIPS network (Fig. 1A, right-hand side) shows a connection between Cdc45 and MCM domains that has also been observed in protein interactions. MCM is also linked to the DNA-binding BAH domain (bromo-adjacent homology) that plays a role in linking DNA methylation, replication and transcriptional regulation (Callebaut et al., 1999). Hence, there is a clear functional coupling between these domains that can be translated back to the protein level to identify putative protein interactions in different organisms.

The predictive use of these networks can be illustrated, e.g. by the correlation between a cell division GTP-binding domain (GTP_Cell_Div, IPR000383) and a domain of unknown function (DUF258, IPR004881) in the MIPS network (Fig. 1A, centre). Early investigation into the function of a conserved family of proteins with DUF258 domain suggest a role of GTPase activity possibly linked to translation control (Daigle et al., 2002). Hence, this correlation can be a starting point for the future exploration of DUF258-type protein family.

The Ezrin/radixin/moesin domain (ERM, IPR000798) is linked to myosin domains in the MIPS network (Fig. 1A, bottom middle). ERM proteins crosslink actin filaments with the plasma membrane (Tsuchida et al., 1997) and myosin proteins drive movement along actin filaments in an ATP-dependent manner during muscle contraction, cytokinesis and vesicle transport (Lodish et al., 2000). The direct association between ERM-type proteins and myosin is not known, but it is clear that both domains are involved in actin-specific cytoskeletal activities.

Another association was identified between Skp1 (IPR001232) and Cullin (IPR001373) domains in the YPD, MIPS and the HMS-PCI networks. Both domains are involved in the E3 complex (also known as the Skp1-Cdc53/cullin-F box, SCF), which is responsible for ubiquitination of proteins targeted for degradation (Tyers and Willems, 1999). Other notable correlations involve transcription initiation that includes the Zinc finger, histone acetyltransferases and DNA binding domains (Fig. 1B).

Owing to the large number of false positive proteins identified in the high-throughput generated complexes, it is hard to systematically discern the biologically relevant domain associations from other fortuitous relationships. However, the integration of additional protein interaction information to identify supported domain correlation facilitates the identification of biologically meaningful correlations. Among the distinct domain clusters in the TAP dataset (Fig. 2A) is a group of small GTPase domains closely related to Ras-mediated signal transduction small GTPases that were also found correlated in the HMS-PCI set. This cluster is supported by protein interaction data from the literature and in a number of independent domain pairs by Sprinzak et al. (Sprinzak and Margalit, 2001), which identified domain pairing from binary protein interactions. In the HMS-PCI network (Fig. 2B), we identified the SH3 domain linked to two transport domains—General Substrate Transporter (IPR005828) (this correlation is supported by a number of Y2H experiments) and VPS10 (IPR006581), which contains an SH3-binding region; consistent with the established role of SH3 in vesicular trafficking (McPherson, 1999).

**CONCLUSIONS**

The discovery of novel protein interactions continues at a rapid pace with the recent addition of the *Drosophila melanogaster* and *Caenorhabditis elegans* interactions (Giot et al., 2003; Li et al., 2004), and more are expected in the near future. The need to assess critically the quality of the data has remained largely an unsolved problem and requires the use of computational methods. We introduced one such approach that focuses on the functional content represented within protein complexes. We have demonstrated that protein complexes can be transformed into a network of domain correlations that capture the functional coupling of domains. These networks can then be used as a starting point for the identification of potential protein associations either in molecular complexes or pathways, as well as, for the identification of specific experimental biases within high-throughput data. The key to the successful identification of biologically relevant information is the integration of multiple sources of data produced by independent complementary interaction experiments in conjunction with different analysis tools.

**ACKNOWLEDGEMENTS**

We thank Ian Donaldson for his insightful comments. Funding from the Canadian Institute of Health Research (CIHR), Genome Canada and Ontario R&D Challenge Fund (ORDCF) supported this study. D.B. was also supported by the University of Toronto Open Fellowship.

**REFERENCES**


